Cost-aware Active Learning for Named Entity Recognition in Clinical Text

Qiang Wei
School of Biomedical Informatics
The University of Texas Health Science Center at Houston
Houston, TX, USA
Qiang.Wei@uth.tmc.edu

Yukun Chen
Pieces Technologies Inc
Dallas, TX, USA
yukun.k.chen@gmail.com

Mandana Salimi
School of Biomedical Informatics
The University of Texas Health Science Center at Houston
Houston, TX, USA
MSalimi@mdanderson.org

Joshua C. Denny
Department of Biomedical Informatics
Vanderbilt University
Nashville, TN
josh.denny@vumc.org

Qiaozhu Mei
School of Information
University of Michigan
Ann Arbor, MI, USA
qmei@umich.edu

Thomas A. Lasko
Department of Biomedical Informatics
Vanderbilt University
Nashville, TN
tom.lasko@vumc.org

Qingxia Chen
Biostatistics
Vanderbilt University
Nashville, TN
cindy.chen@vumc.org

Amy Franklin
School of Biomedical Informatics
The University of Texas Health Science Center at Houston
Houston, TX, USA
Amy.Franklin@uth.tmc.edu

Trevor Cohen
Department of Biomedical Informatics and Medical Education
University of Washington
Seattle, WA
cohenta@uw.edu

Stephen Wu
School of Biomedical Informatics
The University of Texas Health Science Center at Houston
Houston, TX, USA
Stephen.T.Wu@uth.tmc.edu

Hua Xu
School of Biomedical Informatics
The University of Texas Health Science Center at Houston
Houston, TX, USA
Hua.Xu@uth.tmc.edu

ABSTRACT

Active Learning (AL) attempts to reduce annotation cost (i.e., time) and improve model performance by iteratively selecting the most informative examples for annotation. For AL work in the medical domain, most approaches tacitly (and unrealistically) assume that the cost for annotating each sample is identical and many of them evaluated proposed methods in simulation settings, which do not reflect the actual performance of AL in real-time annotation. In this work we designed a novel, cost-aware AL algorithm (Cost-CAUSE) for the task of named entity recognition (NER) in clinical text, which integrates both lexical and syntactic features to build models to estimate annotation time for a given sample. Using the 2010 i2b2/VA dataset, we recruited 9 users and conducted a user study to compare Cost-CAUSE with passive learning in a real-time NER annotation task. Our results show that Cost-CAUSE outperformed passive learning on the ALC score and reduced annotation time by 20.5-30.2%, demonstrating the great potential of AL in clinical NER.

KEYWORDS

Active learning, Named entity recognition, Natural language processing, Electronic health records

1 Introduction

Supervised machine learning (ML) models have achieved state-of-the-art performance across a range of clinical Natural Language Processing (NLP) tasks[6,13], but statistical NLP systems often require large numbers of annotated samples in order to build high performance ML models. Constructing large-scale, high-quality corpora is time consuming and costly, particularly in the medical domain where corpus-building often requires manual annotation by domain experts. Therefore, methods that can help build high-performance ML models but require fewer annotations are highly desirable in clinical NLP.

Active learning (AL) systems attempt to prioritize more informative samples for annotation during an iterative training process to reduce time consuming. Although some studies demonstrated the potential of AL, they were conducted in simulated environments, and assumed that annotation costs for each sample were identical[1–3,9]. However, in reality, annotation cost (i.e. the time required by an annotator) can be very different from one sample to another and from one user to another user. Cost-conscious AL methods take consideration of annotation cost together with sample informativeness, by integrating a cost (i.e., time) model to estimate annotation time for unlabeled data[7][11], and have shown improved performance in real-time annotation[10]. However, in the medical domain, few studies have investigated cost-conscious AL methods and none was evaluated in in the context of real-time user studies. Consequently these studies do not address important aspects of the complexity of annotation in production environments, such as differences in annotation quality across users and the element of user fatigue, both of which affect the estimated performance of AL. In this study we developed a

novel cost-aware AL approach for clinical NER and showed that it outperformed passive learning in real-time annotation. The contribution of the current work can be summarized as follows:
(1) We develop a new AL algorithm for clinical NER (called Cost-CAUSE), which extends CAUSE (Clustering And Uncertainty Sampling Engine) [1] to weigh estimated annotation costs (by considering lexical and syntactic complexity of sentences) against informativeness in one model.
(2) We conduct the first user study to evaluate our AL system and a passive learning system at a real-time setting for clinical NER annotation.

2 Related Work

2.1 AL in simulation

Many studies of AL were done in simulation settings, which assume annotation cost for each sample is same. Settles and Craven conducted a large scale general-domain evaluation of multiple AL methods, with a number of approaches giving relatively robust performance[9]. Chen et al. demonstrated that AL outperformed random sampling for a simulated clinical NER task[2]. More recently, Kholghi et al. also applied AL to simulated clinical NER tasks and showed that AL could reach the same accuracy as random sampling, with only 54% (i2b2/VA 2010) and 76% (ShARe/CLEF 2013) of the total number of concepts in the training data[3].

2.2 AL and Real-Annotation Time

Despite these studies demonstrated the potential of AL, they were conducted in simulated environments, and assumed that annotation cost for each sample was identical, which is not true in reality. To address this issue, Settles et al.[10] collected several corpora along with sample-level annotation times to evaluate real-world AL performance, and found that observed costs are highly variable across instances. Chen et al.[1] developed an AL annotation system to sample sentences for users, concluding on the basis of user studies that cost-agnostic AL approaches may perform no better than random sampling on the measurement of annotation time, but improved learning curves are achievable if the cost variables can be appropriately taken into account. Nevertheless, Kholghi et al.[4] recruited four users to compare various AL methods with random sampling. The AL methods in their tests reduced annotation time by 28% compared with random sampling.

2.3 Cost-Conscious AL

To address the issue that AL may not actually reduce the annotation time in reality, a number of studies proposed methods to model the annotation cost of a sample, and balance that cost against informativeness. Haertel et al.[7] presented a practical cost-conscious AL approach motivated by the business concept of return on investment (ROI), and showed a 73% reduction in hourly cost as compared with random sampling on a POS tagging task. Tomanek et al.[11] summarized and compared several methods that incorporated cost variables into AL and found that using the ratio between informativeness and cost was an effective way to incorporate a cost variable into AL. However, how to build models to estimate the actual cost of sample annotation is still an open question and it could be very different depending on the NLP task. In this study of AL in clinical NER, a sample is a sentence; therefore, we will investigate models to estimate cost (time) of annotating one sentence.

3 Materials and Methods

3.1 Dataset

The dataset used in the study came from 2010 i2b2/VA challenge, preserving the original training and test splits of 349 clinical documents (20,423 unique sentences) and 477 clinical documents (29,789 unique sentences)[12]. Three types of medical entities were annotated in each sentence: "problem", "treatment", and "test".

3.2 Cost-CAUSE

We propose Cost-CAUSE as an approach to identify more-informative, less-costly sentences. While we follow the CAUSE[1] query strategy to select unlabeled sentences for annotation, we score sentences using the ratio: Informativeness(s)/Cost(s) between the informativeness of a sentence s and its estimated annotation time, similar to other cost-conscious approaches[8,9]. Cost-CAUSE is encapsulated by the following pseudocode (Figure 1). The ranked sentence set *S* maintains a balanced distribution across topics while selecting high IPC samples, and top sentences in *S* will be used for annotating. here the *Informativeness* for sentence *s* is entropy of words in the *s*.

1. Cluster sentences *s* into groups *g* according to their topics.
2. Calculate informativeness per cost (IPC) for each sentence: IPC(s)=*Informativeness(s)/Cost(s)*
3. Calculate averaged IPC for each group g_i: Avg IPC $(g_i) = \frac{\sum_{s \in g_i} IPC(s)}{\#(sentences\ in\ g_i)}$
4. Ranked group list → G: $g_1, g_2, \ldots, g_n$
5. For g_i in G:
   select sentence *s* with highest IPC in g_i
   put *s* into ranked sentence set *S*
   remove *s* from g_i

Figure 1: Cost-CAUSE algorithm.

3.3 Cost Model

Motivated by psycholinguistic literature[5,14], we developed a cost model to estimate annotation cost for one sentence based on features selected to capture the basic characteristics, lexical complexity, and syntactic complexity of the sentence. The cost model is given by the following formula:

$$Cost(s) = c_0 + \sum_i c_i f_i(s) \quad (1)$$

where $f_i(s)$ is the value of feature *i* for sentence *s*, and coefficients c_i are parameters learned during training. The coefficients in formula was learned using least squares.

Figure 2 shows all the features used in the study. The Count features reflect the characteristics of sentences such as their length. The POS Tag Entropy feature is based on the corpus-derived probabilities of POS bigrams. The cumulative inverse document frequency (IDF, with sentences as "documents") is used to measure the lexical complexity on the assumption that infrequently encountered terms may take longer to process.

Sentence	***MRI*** by report showed ***bilateral rotator cuff repairs*** and he was admitted for ***repair of the left rotator cuff*** .				
Categories	Count			Lexicon	Syntactic
Feature	Number of words (NOW)	Number of entities (NOE)	Number of entity words (NOEW)	Inverse Document Freq. (IDF)	Entropy of POS tag (EOP)
Value	20	3	11	35.36	2.28

Figure 2: Features in the cost model.

3.4 User Study

We conducted a user study to compare overall NER annotation times using Cost-CAUSE vs. random sampling. The participants were recruited in the University of Texas Health Science Center at Houston, and met the following criteria: 1) they were medical or nursing students; 2) they had experience working with clinical notes written in English. The actual training and evaluation were conducted in three phases, designed for consistency in annotation skill level and environment.

Phase I. 20 participants took basic NER annotation training, and were tested to characterize their level of accuracy.

Phase II. All 20 participants entered phase 2, but 8 participants discontinued the user study for personal reasons. Data used in phase II came from the training set of i2b2/VA. Participants took a further training and practice. Finally, they took a 1-hour test to determine their eligibilities for Phase III, and data from the test were used for fit parameters of a cost model for each participant.

Phase III. Phase III was conducted in two days. For each day, participants reviewed their annotation from the Phase II test for a half hour to warm up. Then, they took the annotation test using our annotation system with the i2b2 2010 test dataset for 120 minutes. The test was in three 40-minutes sessions and there was a 15-minute break between any two sessions. A total of 10 medical experts completed phase 3, but one participant was subsequently removed from further analysis due to lower annotation quality. Dataset for evaluating models trained from users' annotation in this phrase came from the original training set of the i2b2/VA.

The learning curves that plot F-measures vs. estimated annotation time were generated to visualize the performance of different methods. The area under the learning curve (ALC) was used to compare the performance of AL or PL methods.

4 Results

Performance of AL and PL. AL (Cost-CAUSE) outperformed PL (random selection) for eight of nine users in ALC scores (Table 1 and Figure 3), and the score for AL was significantly larger than that for PL (Wilcoxon signed-rank test, $p < 0.01$). At 120 minutes, the ML model trained on AL sentences had a better performance than that from PL sentences in seven of nine users. To test whether AL is significantly different from PL in terms of performance of the ML model as captured by the learning curves, we performed a Wilcoxon signed-rank test for each user, and AL significantly outperformed PL in terms of ALC for six of nine users ($p < 10^{-3}$).

Annotation Performance. The annotation quality F-measure was estimated by comparing user annotations to the reference standard. While users maintained at least 0.70 F-measure on annotation quality, there was an observable difference between AL and PL (0.748 for AL and 0.798 for PL on average; a median of 0.74 for AL and 0.79 for PL). Annotation qualities of three users for AL sentences were much lower than PL (user 1, 3 and 4, ~0.08 lower). Users spent a longer time annotating words in AL sentences (33.01 - 73.07 vs. 40.47 - 92.22 words/minute) and annotated fewer AL sentences within 120 minutes. AL sentences were slightly longer (12.44 vs. 11.38 words/sentence on average), contained more entities (2.14 vs. 1.39 entities/sentences on average), had a higher entity density (0.34 vs. 0.26 on average), and thus were perhaps more difficult for users.

Table 1: ALC scores, F-measures at the end of 120-minute annotation, and the statistical test P-values of AL and PL. Best performance across models for a user is in boldface.

Users	ALC scores		F-measures at 120 minutes		P-values based on Wilcoxon signed-rank test
	PL	AL	PL	AL	
User1	0.633	**0.637**	**0.696**	0.695	$9.7\text{x}10^{-2}$
User2	0.574	**0.575**	0.659	**0.671**	$7.2\text{x}10^{-3}$
User3	0.608	**0.628**	0.683	**0.690**	$3.0\text{x}10^{-5}$
User4	**0.615**	0.609	**0.692**	0.680	$5.6\text{x}10^{-3}$
User5	0.619	**0.642**	0.707	**0.717**	$1.8\text{x}10^{-5}$
User6	0.580	**0.610**	0.674	**0.691**	$3.9\text{x}10^{-4}$
User7	0.521	**0.580**	0.624	**0.671**	$1.8\text{x}10^{-5}$
User8	0.599	**0.613**	0.673	**0.691**	$2.7\text{x}10^{-5}$
User9	0.606	**0.629**	0.683	**0.693**	$1.8\text{x}10^{-5}$
Mean	0.595	**0.632**	0.677	**0.689**	

There was a decrease in annotation quality for AL from 40 minutes (Figure 3, blue dashed lines), where annotation quality for AL clearly falls below PL. This may be because sentences selected by AL become progressively harder to annotate as sentences must be more atypical to qualify as "informative" as the model evolves.

Annotation Effort Saved by Cost-CAUSE. Consider a complementary measurement of users' annotation effort: how much annotation is necessary (in minutes, number of sentences, and number of words) to reach a target performance F-measure of 0.67 (a higher threshold will result in that most of the PL models couldn't reach the target performance)? For users 1 and 4, the AL model took more time to reach target performance than PL. For another two users 2 and 7, the PL model never reached the target performance at the end of 120 minutes, while the AL model did. It

needed 74.9 minutes for AL and 86.8 minutes for PL to reach target performance on average. For the remaining five users, AL reduced the annotation time to reach the target performance by 20.5% - 30.2%, and reduced the number of sentences and words annotated at target performance by 43% - 49.4% and 37.6% - 44.4% respectively. Interestingly, although Cost-CAUSE did not reduce annotation time for User 1 and User 4, it did reduce the number of annotated sentences. However, as we have argued previously, annotation time is a more important measure of performance for practical purposes.

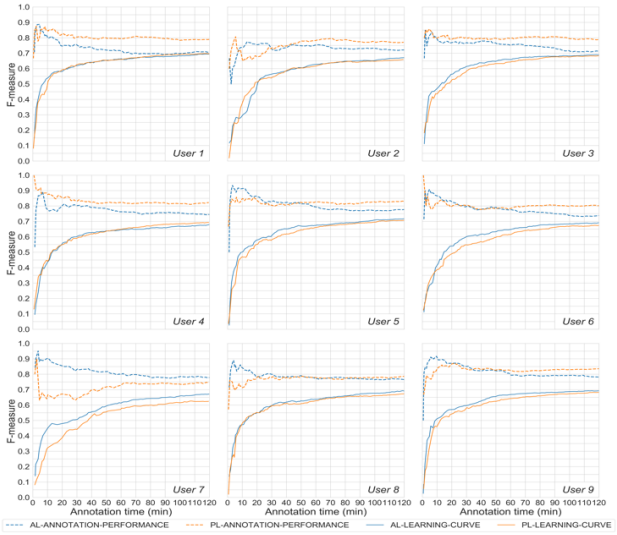

Figure 3: Learning curves and annotation performance for the 9 users. Dashed lines represent annotation quality and solid lines represent the learning curves. The orange and blue color represent PL and AL respectively.

Factors Related with Performance of the AL method. AL failed to outperformed the PL for user 1 and 4, which may be because their annotation quality for AL was much lower than their annotation quality for PL. Another possible reason may be it's more difficult for our cost model to estimate annotation cost for the two users compared with other users.

Limitations of this work: There is still some limitations in our work. Our cost model only takes textual features of sentences into consideration and doesn't make use of users' the background, reading speed and other characteristics. So in the future an interesting research direction may be to develop more sophisticated annotation time models that consider these features.

5 CONCLUSION

In this study, we presented a cost model to predict annotation time, which was then integrated into the new AL querying algorithm (Cost-CAUSE). Cost-CAUSE was shown to save 20.5-30.2% annotation time for 9 users in a two-hour annotation experiment using the i2b2 2010 dataset. These results demonstrate the importance of considering, and compensating for, the cost of sentence annotation in AL-based clinical NER.

ACKNOWLEDGEMENTS

This work was supported by National Library of Medicine grant number 2R01LM010681-05.

COMPETING INTERESTS

Dr. Xu and The University of Texas Health Science Center at Houston have research-related financial interests in Melax Technologies, Inc.

REFERENCES

[1] Yukun Chen, Thomas A. Lask, Qiaozhu Mei, Qingxia Chen, Sungrim Moon, Jingqi Wang, Ky Nguyen, Tolulola Dawodu, Trevor Cohen, Joshua C. Denny, and Hua Xu. 2016. An Active Learning-enabled Annotation System for Clinical Named Entity Recognition. *BMC Med. Inform. Decis. Mak.* 17, Suppl 2 (July 2016), 82. DOI:https://doi.org/10.1186/s12911-017-0466-9

[2] Yukun Chen, Thomas A Lasko, Qiaozhu Mei, Joshua C Denny, and Hua Xu. 2015. A study of active learning methods for named entity recognition in clinical text. *J. Biomed. Inform.* 58, (December 2015), 11–8. DOI:https://doi.org/10.1016/j.jbi.2015.09.010

[3] Mahnoosh Kholghi, Laurianne Sitbon, Guido Zuccon, and Anthony Nguyen. 2015. Active learning: a step towards automating medical concept extraction. *J. Am. Med. Inform. Assoc.* (August 2015). DOI:https://doi.org/10.1093/jamia/ocv069

[4] Mahnoosh Kholghi, Laurianne Sitbon, Guido Zuccon, and Anthony Nguyen. 2017. Active learning reduces annotation time for clinical concept extraction. *Int. J. Med. Inform.* 106, (October 2017), 25–31. DOI:https://doi.org/10.1016/j.ijmedinf.2017.08.001

[5] Gina R Kuperberg and T Florian Jaeger. 2016. What do we mean by prediction in language comprehension? *Lang. Cogn. Neurosci.* 31, 1 (2016), 32–59. DOI:https://doi.org/10.1080/23273798.2015.1102299

[6] Feifan Liu, Jinying Chen, Abhyuday Jagannatha, and Hong Yu. 2016. Learning for Biomedical Information Extraction: Methodological Review of Recent Advances. (June 2016). Retrieved March 16, 2017 from http://arxiv.org/abs/1606.07993

[7] James L. Carroll Robbie A. Haertel, Eric K. Ringger. 2008. Return on Investment for Active Learning. In *In Proceedings of the Neural Information Processing Systems Workshop on Cost Sensitive Learning2008*. Retrieved April 30, 2016 from http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.148.9408

[8] Burr Settles. 2009. Active Learning Literature Survey. In *Computer Sciences Technical Report 1648*. Retrieved January 23, 2018 from http://burrsettles.com/pub/settles.activelearning.pdf

[9] Burr Settles and Mark Craven. 2008. An analysis of active learning strategies for sequence labeling tasks. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 1070–1079. Retrieved February 10, 2016 from http://dl.acm.org/citation.cfm?id=1613715.1613855

[10] Burr Settles, Mark Craven, and Lewis Friedland. 2008. Active Learning with Real Annotation Costs. In *Proceedings of the NIPS Workshop on Cost-Sensitive Learning*, 1–10. Retrieved April 30, 2016 from http://burrsettles.com/pub/settles.nips08ws.pdf

[11] Katrin Tomanek and Udo Hahn. 2010. A comparison of models for cost-sensitive active learning. (August 2010), 1247–1255. Retrieved April 4, 2016 from http://dl.acm.org/citation.cfm?id=1944566.1944709

[12] Özlem Uzuner, Brett R South, Shuying Shen, and Scott L DuVall. 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text. *J. Am. Med. Inform. Assoc.* 18, 5 , 552–6. DOI:https://doi.org/10.1136/amiajnl-2011-000203

[13] Yanshan Wang, Liwei Wang, Majid Rastegar-Mojarad, Sungrim Moon, Feichen Shen, Naveed Afzal, Sijia Liu, Yuqun Zeng, Saeed Mehrabi, Sunghwan Sohn, and Hongfang Liu. 2018. Clinical information extraction applications: A literature review. *J. Biomed. Inform.* 77, (January 2018), 34–49. DOI:https://doi.org/10.1016/J.JBI.2017.11.011

[14] Stephen Wu, Asaf Bachrach, Carlos Cardenas, and William Schuler. 2010. Complexity Metrics in an Incremental Right-Corner Parser. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, 1189–1198. Retrieved October 19, 2018 from http://www.aclweb.org/anthology/P10-1121